HAL
archives-ouvertes.fr

Tunicates and not cephalochordates are the closest living relatives of vertebrates.

Frédéric Delsuc, Henner Brinkmann, Daniel Chourrout, Hervé Philippe

▶ **To cite this version:**

Frédéric Delsuc, Henner Brinkmann, Daniel Chourrout, Hervé Philippe. Tunicates and not cephalochordates are the closest living relatives of vertebrates.. Nature, Nature Publishing Group, 2006, 439 (7079), pp.965-8. <10.1038/nature04336>. <halsde-00315436>

HAL Id: halsde-00315436
https://hal.archives-ouvertes.fr/halsde-00315436

Submitted on 28 Aug 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tunicates and not cephalochordates are the closest living relatives of vertebrates

Frédéric Delsuc[1*], Henner Brinkmann[1], Daniel Chourrout[2] & Hervé Philippe[1]

[1] *Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada*

[2] *Sars Centre for Marine Molecular Biology, Bergen High Technology Centre, University of Bergen, Thormøhlensgaten 55, 5008 Bergen, Norway*

* Present address: Laboratoire de Paléontologie, Phylogénie et Paléobiologie, Institut des Sciences de l'Evolution, UMR 5554-CNRS, Université Montpellier II, France.

Tunicates or urochordates (appendicularians, salps, and sea squirts), cephalochordates (lancelets) and vertebrates (including lamprey and hagfish) constitute the three extant groups of chordate animals. Traditionally, cephalochordates are considered as the closest living relatives of vertebrates with tunicates representing the earliest chordate lineage[1,2]. This view is mainly justified by overall morphological similarities and an increased complexity in cephalochordates and vertebrates relative to tunicates[2]. Despite their critical importance for understanding the origins of vertebrates[3], phylogenetic studies of chordate relationships have provided equivocal results[4-7]. Here, taking advantage of the genome sequencing of the appendicularian *Oikopleura dioica*, we assembled a phylogenomic dataset of 146 nuclear genes (33,800 unambiguously aligned amino acids) from 14 deuterostomes and 24 other slowly evolving species as an outgroup. We show that phylogenetic analyses of this dataset provide compelling evidence that tunicates, and not cephalochordates, represent the closest living relatives of

vertebrates. Moreover, chordate monophyly remains uncertain since cephalochordates, albeit with a non-significant statistical support, surprisingly grouped with echinoderms, a hypothesis that needs to be tested with additional data. This new phylogenetic scheme prompts a reappraisal of both morphological and palaeontological data and has important implications for the interpretation of developmental and genomic studies in which tunicates and cephalochordates are used as model animals.

The introduction of molecular data into classical systematics has already put to test a number of evolutionary hypotheses through the analysis of individual genes such as ribosomal RNA (rRNA). However, phylogenies reconstructed from single or a small number of genes are hampered by stochastic effects limiting the statistical significance of the results. The genomic era is now providing the opportunity for phylogenetics to resolve a number of outstanding evolutionary questions through an increase of resolving power[8]. This applies to the origin and early evolution of vertebrates, a fundamental evolutionary question that has been revived by recent advances in molecular and developmental biology as well as new fossil discoveries[3]. The understanding of these events has to be considered in the context of chordate phylogeny where the traditional textbook view considers cephalochordates as the closest living relatives of vertebrates (a group named Euchordata), to the exclusion of the morphologically more distinct tunicates[2]. Although almost universally accepted, this classical picture is supported by only a limited number of morphological features that are far from being unambiguous. For example, the presence of metameric segmentation[1] used to link cephalochordates and vertebrates might in fact be considered as an ancestral feature of deuterostomes[9]. The classical view (Euchordata) has also found some support in molecular studies of rRNA genes[5]. However, a competing hypothesis grouping tunicates and vertebrates into a clade named Olfactores[10] was recovered in cladistic analyses of combined rRNA and morphology[4] and suggested by the structure of cadherin genes[11]. However, the

statistical significance of these apparently conflicting results was limited by the relatively few characters considered.

Recently, two multigene studies based on nuclear proteins have provided some support for Olfactores[6,7]. However, the extremely limited chordate species sampling considered in these studies prevented drawing any firm conclusions given its potentially deleterious effect on phylogenetic inference[8]. We have therefore extended the Philippe *et al.* dataset[7] of 146 genes from four to 13 chordates including one cephalochordate, four tunicates, and eight vertebrates with the notable inclusion of the early-branching agnaths (hagfish and lamprey). Within tunicates, the incorporation of *O. dioica* is particularly important since it belongs to appendicularians (or larvaceans), which are morphologically and molecularly very divergent from the ascidians previously included.

Phylogenetic analyses of this multigene dataset using maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference all converged to the same topology (Fig. 1). The statistical support was maximal in Bayesian analyses, where all nodes received a posterior probability of 1.0. All non controversial groups (choanoflagellates, cnidarians, molluscs, arthropods, tunicates and vertebrates) were recovered with strong bootstrap support ($BP_{MP-ML} > 95\%$), as were also metazoans, bilaterians, protostomes, and lophotrochozoans (Fig. 1). Weak statistical support ($BP_{MP-ML} < 72\%$) was only observed for some relationships within insects and bivalves. Within vertebrates, our results strongly support the controversial monophyly of cyclostomes (lamprey and hagfish)[2]. Also, both MP and ML provided reasonable support for the monophyly of deuterostomes ($BP_{MP-ML} = 87\%$-93%). However, within deuterostomes, chordates appeared not to be monophyletic as cephalochordates grouped with echinoderms, albeit with moderate ML bootstrap support ($BP_{MP-ML} = 97\%$-89%). By contrast, whereas MP moderately supported the grouping of tunicates and vertebrates (90%), the more

accurate ML method[12] provided unambiguous bootstrap support (100%) for Olfactores (Fig. 1).

To further test the stability of phylogenetic relationships within deuterostomes, we evaluated in a likelihood framework the 15 rooted topologies corresponding to all possibilities of connecting the four major groups under study (echinoderms, cephalochordates, tunicates and vertebrates). 13 alternatives to the ML topology were significantly rejected at the 5% confidence level by all statistical tests (Table 1). Only the topology where chordates are monophyletic with cephalochordates joining the tunicates plus vertebrate clade was not rejected (Table 1). The traditional hypothesis of euchordate monophyly was ranked only 4th in terms of log-likelihood, after the alternative in which cephalochordates emerge before echinoderms. These two topologies appeared significantly worse that the ML tree of Fig. 1, even for the conservative SH test[13].

Our results therefore indicate a strong phylogenetic affinity between tunicates and vertebrates to the exclusion of cephalochordates. However, obtaining high statistical support for a given topology does not necessarily indicate that the phylogenetic inference is correct. Indeed, the phylogenetic analysis of large-scale datasets requires particular attention to potential systematic biases associated, for instance, with differences in evolutionary rates among species, compositional biases and heterotachy[8]. In particular, a long-branch attraction (LBA) artefact[14] may potentially occur since tunicates include fast (*Ciona ssp.*) and very fast (*O. dioica*) evolving species (Fig. 1). A high evolutionary rate of tunicate genes was already noticed in rRNA genes[5] and in complete mitochondrial genomes[15]. Our results confirm these observations for a large number of nuclear genes. As fast evolutionary rates are also often associated with compositional bias or with heterotachy, it is a necessary first step to exclude the

possibility that the observed grouping of tunicates with vertebrates results from a tree reconstruction artefact.

The most obvious potential artefact, LBA, predicts that the fast evolving tunicates would be attracted towards the outgroup, and not by the slowly evolving vertebrates. This would produce a topology compatible with the classical hypothesis of chordate evolution where the slow evolving cephalochordates and vertebrates group together. This prediction is perfectly congruent with the lower support for Olfactores observed with MP (90%), a method known to be more sensitive to LBA than probabilistic methods[14]. Indeed, when *O. dioica* was used as the single representative of tunicates, MP unambiguously supported (BP = 100) an aberrant position for this group which emerged before cnidarians, disrupting the monophyly of bilaterians (Fig. S1). By contrast, the less sensitive ML method recovered Olfactores, albeit with decreased bootstrap support (BP = 84) (Fig. S2). Therefore, despite its extreme evolutionary rate, *O. dioica* retained enough phylogenetic signal for its position to be recovered with ML. This demonstrates that LBA is not responsible for the inferred grouping of tunicates and vertebrates, and represents a strong argument in favour of the authenticity of Olfactores. In addition, neither compositional bias nor heterotachy significantly influenced phylogenomic inference with our dataset (see Supplementary Information). In fact, the compositional effect would act against Olfactores since vertebrates and the amphioxus shared similar amino acid compositions (Fig. S3). In conclusion, the strongly supported monophyly of Olfactores cannot be explained by any kind of identifiable systematic biases (LBA, compositional bias, and heterotachy) and therefore constitutes the best current hypothesis for chordate phylogeny.

The monophyly of deuterostomes remained moderately supported in our phylogenomic analyses (Fig. 1). Also, the monophyly of chordates is not found in the ML tree, but is the only alternative not significantly rejected by likelihood-based

statistical tests (Table 1). A unique origin of chordates and their distinctive features such as notochord and hollow nerve cord cannot be excluded. Our results nevertheless favoured the intriguing possibility of a sister-group relationship between cephalochordates and echinoderms that seems robust to analyses aimed at avoiding compositional bias and heterotachy (see Supplementary Information). Such a relationship has also been inferred from mitochondrial genomes[16], but it lacks significant statistical support in both nuclear and mitochondrial analyses.

Although seemingly heretic, the grouping of echinoderms and cephalochordates constitutes an interesting working hypothesis. A similar situation was encountered a few years ago for the recently established sister-group relationship of echinoderms and hemichordates (Ambulacraria). The Ambulacraria hypothesis led to a re-evaluation of morphological character evolution with the presence of pharyngeal slits being interpreted as an ancestral feature of the deuterostome ancestor[9]. Similarly with the present case, a close relationship between echinoderms and cephalochordates would imply that a dorsal nerve chord was already present in the last common deuterostome ancestor and subsequently evolved into derived nervous systems in both hemichordates and echinoderms. Such a scenario seems *a priori* possible given the difficulties encountered in polarising morpho-anatomical characters in both extant[9] and fossil[17] deuterostomes. However, a definitive conclusion will only be achieved through the phylogenetic analysis of more genes combined with an increased taxon sampling including the enigmatic xenoturbellidans, hemichordates, and a greater diversity of echinoderms. Nonetheless, the strong support obtained for Olfactores will likely not be affected, as these additional taxa are considered to be on the echinoderm side of the deuterostome tree[5,18]. This prediction is supported by the observation that removing the sea-urchin from our dataset has virtually no effect (Fig. S4).

Despite this remaining uncertainty, our new phylogenetic hypothesis implies a serious re-evaluation of fundamental aspects of deuterostome evolution. The nature of their last common ancestor has been most extensively addressed from the paleontological point of view[3]. However, extant deuterostome lineages are morphologically so distinct that possible stem-group representatives found in the fossil record are difficult to recognise[17]. A sister-group relationship of tunicates and vertebrates to the exclusion of cephalochordates is compatible with the controversial calcichordate theory of chordate origins proposed by Jefferies[19]. However, it does not mean that this evolutionary scenario based on the functional reconstruction of unusual fossils with calcite skeletons (cornutes and mitrates) and their interpretation as stem-group chordates[19] is necessary true. In fact, Jefferies[10] coined the name Olfactores on the basis of the presence of a homologous olfactory apparatus in fossils proposed to be precursors of tunicates and vertebrates. However, the phylogenetic position of cornutes and mitrates is still highly debated with the majority advocating for echinoderm affinities of these controversial fossils[20,21]. At any rate, the present molecular evidence for a monophyletic group of tunicates and vertebrates might help to polarize morphological characters of basal deuterostome fossils, thereby leading to a better understanding of early deuterostome evolution.

Our results also prompt a reinterpretation of morphological data in deuterostome phylogeny. In particular, a close proximity between tunicates and vertebrates suggests that the presence of metameric segmentation classically used to unify cephalochordates and vertebrates might be considered as an ancestral feature that underwent a secondary reduction in tunicates[9]. More generally, this new phylogenetic picture is in agreement with an alternative hypothesis for chordate evolution based on a recent homology analysis of morphological structures in hemichordates and chordates[22]. This unorthodox view proposes that cephalochordates have retained many ancestral characters that have been secondarily lost in the morphologically more derived tunicates and reveals 13

putative synapomorphies uniting tunicates and vertebrates to the exclusion of cephalochordates[22]. The monophyly of Olfactores invalidates the traditional textbook representation of chordate, and even deuterostome evolution[9], as a steady increase towards complexity culminating in the highly specialized brain of vertebrates. This anthropocentric interpretation is perhaps best reflected by the terms “Euchordata” (i.e. “true chordates”) or “chordates with a brain” used to designate the grouping of cephalochordates and vertebrates[2]. Tunicates should therefore no longer be considered as “primitive” but rather as derived chordates with highly specialized lifestyles and developmental modes.

From the developmental point of view, our phylogenetic results help to understand the origin of the major evolutionary novelty constituted by the neural crest. This vertebrate innovation can be traced back to the origins of the chordate lineage since “latent homologues” of neural crest cells have been identified in both cephalochordates and tunicates[23]. However, evidence for migratory neural crest cells has so far only been reported in tunicates[24], whereas their existence is still unproven in amphioxus. In light of the Olfactores hypothesis, these migratory cells may well have evolved in the last common ancestor of tunicates and vertebrates, after the divergence from cephalochordates, with these evolutionary precursor cells latter giving birth to the neural crest along the vertebrate lineage[24].

The newly proposed deuterostome phylogeny strengthens the view that tunicates and cephalochordates represent complementary models for studying the origin of the vertebrate developmental program. Indeed, tunicates are phylogenetically closer to vertebrates but are morphologically and molecularly highly derived with a trend towards genomic simplification[25,26], whereas the more distantly related cephalochordates might have retained more ancestral characters[27]. The comparative analysis of available tunicate and vertebrate genomes with the upcoming amphioxus and

sea-urchin genome sequences will be particularly valuable for understanding the evolution of new gene systems and structures involved in early vertebrate development.

Methods

Data assembly. We built upon a phylogenomic dataset consisting of 146 nuclear genes previously assembled to study animal phylogeny[7]. This dataset was updated using the same protocol (see Supplementary Information) with new sequences publicly available from the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/) and the EST Database (http://www.ncbi.nlm.nih.gov/dbEST/) of GenBank at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). Sequences from the appendicularian were generated by the *Oikopleura dioica* genome project (http://www.genoscope.cns.fr/externe/English/Projets/Projet_HG/HG.html).

As previously demonstrated[7,8], taxon sampling has a major impact in phylogenomic studies. As an outgroup to the 14 available deuterostomes, we therefore selected the slowest evolving taxa among available protostomes and fungi in order to reduce the potential impact of long-branch attraction[14]. Furthermore, we also incorporated all available cnidarians and choanoflagellates allowing to efficiently break the long-branch leading to the distantly related fungal outgroup.

Phylogenetic analyses. Multiple methods using different optimality criteria and algorithms were used to analyse our phylogenomic dataset. Weighted MP heuristic searches were conducted using PAUP[28] with 10 random additions of species and TBR branch swapping. MP bootstrap percentages were obtained after 1,000 replications using the same heuristic search strategy. Given the computational difficulties involved in conducting ML searches for such a large dataset[7], ML analyses were conducted with

different algorithms (see Supplementary Information for details). ML bootstrap percentages were obtained after 100 replications. Bayesian phylogenetic inferences were also conducted using parallel computing (see Supplementary Information for details).

Likelihood-based tests of alternative topologies were calculated using CONSEL[13]. ML branch lengths of alternative topologies were first inferred assuming a concatenated WAG+F+Γ_4 model using TREE-PUZZLE[29], site-wise log-likelihood values were then computed with CODEML[30] and p-values of the different likelihood-based tests were finally calculated with CONSEL.

1. Schaeffer, B. Deuterostome monophyly and phylogeny. *Evol. Biol.* **21**, 179-235 (1987).

2. Rowe, T. in *Assembling the Tree of Life* (eds. Cracraft, J. & Donoghue, M. J.) 384-409 (Oxford University Press, Oxford, 2004).

3. Gee, H. *Before the backbone: views on the origin of the vertebrates* (Chapman & Hall, London, 1996).

4. Zrzavy, J., Mihulka, S., Kepka, P., Bezdek, A. & Tietz, D. Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. *Cladistics* **14**, 249-285 (1998).

5. Winchell, C. J., Sullivan, J., Cameron, C. B., Swalla, B. J. & Mallatt, J. Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. *Mol. Biol. Evol.* **19**, 762-776 (2002).

6. Blair, J. E. & Hedges, S. B. Molecular phylogeny and divergence times of deuterostome animals. *Mol. Biol. Evol.* **22**, 2275-2284 (2005).

7. Philippe, H., Lartillot, N. & Brinkmann, H. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Mol. Biol. Evol.* **22**, 1246-1253 (2005).

8. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6**, 361-375 (2005).

9. Gee, H. in *Major events in early vertebrate evolution: palaeontology, phylogeny, genetics, and development* (ed. Ahlberg, P. E.) 1-14 (Taylor and Francis, London, 2001).

10. Jefferies, R. P. S. in *Biological Asymmetry and Handedness* (eds. Bock, G. R. & Marsh, J.) 94-127 (Wiley, Chichester, 1991).

11. Oda, H., Akiyama-Oda, Y. & Zhang, S. Two classic cadherin-related molecules with no cadherin extracellular repeats in the cephalochordate amphioxus: distinct adhesive specificities and possible involvement in the development of multicell-layered structures. *J. Cell Sci.* **117**, 2757-2767 (2004).

12. Felsenstein, J. *Inferring phylogenies* (Sinauer Associates, Inc., Sunderland, MA, USA, 2004).

13. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246-1247 (2001).

14. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401-410 (1978).

15. Yokobori, S., Oshima, T. & Wada, H. Complete nucleotide sequence of the mitochondrial genome of *Doliolum nationalis* with implications for evolution of urochordates. *Mol. Phylogenet. Evol.* **34**, 273-283 (2005).

16. Ruiz-Trillo, I., Riutort, M., Fourcade, H. M., Baguna, J. & Boore, J. L. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Mol. Phylogenet. Evol.* **33**, 321-332 (2004).

17. Conway Morris, S. The Cambrian "explosion": slow-fuse or megatonnage? *Proc. Natl. Acad. Sci. USA* **97**, 4426-4429 (2000).

18. Bourlat, S. J., Nielsen, C., Lockyer, A. E., Littlewood, D. T. & Telford, M. J. *Xenoturbella* is a deuterostome that eats molluscs. *Nature* **424**, 925-928 (2003).

19. Jefferies, R. P. S. *The ancestry of the vertebrates* (Cambridge University Press, London, 1986).

20. Peterson, K. J. A phylogenetic test of the calcichordate scenario. *Lethaia* **28**, 25-38 (1995).

21. Jefferies, R. P. S. A defence of the calcichordates. *Lethaia* **30**, 1-10 (1997).

22. Ruppert, E. E. Key characters uniting hemichordates and chordates: homologies or homoplasies? *Can. J. Zool.* **83**, 8-23 (2005).

23. Stone, J. R. & Hall, B. K. Latent homologues for the neural crest as an evolutionary novelty. *Evol. Dev.* **6**, 123-129 (2004).

24. Jeffery, W. R., Strickler, A. G. & Yamamoto, Y. Migratory neural crest-like cells form body pigmentation in a urochordate embryo. *Nature* **431**, 696-699 (2004).

25. Seo, H. C. et al. Hox cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature* **431**, 67-71 (2004).

26. Edvardsen, R. B. et al. Remodelling of the homeobox gene complement in the tunicate *Oikopleura dioica*. *Curr. Biol.* **15**, R12-R13 (2005).

27. Holland, L. Z., Laudet, V. & Schubert, M. The chordate amphioxus: an emerging model organism for developmental biology. *Cell. Mol. Life Sci.* **61**, 2290-2308 (2004).

28. Swofford, D. L. *PAUP*: Phylogenetic Analyses Using Parsimony and other methods* (Sinauer, Sunderland, MA, 2000).

29. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502-504 (2002).

30. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555-556 (1997).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank Simon Conway Morris, Richard Jefferies, William Jeffery and two anonymous referees for helpful suggestions, and Nicolas Lartillot and Nicolas Rodrigue for critical readings of early versions of the manuscript. *Oikopleura* genome data have been generated at Génoscope Evry (France) with material and co-funding from the Sars International Centre. We are grateful to Patrick Wincker and the personnel involved at Génoscope. The authors gratefully acknowledge the financial support provided by Génome Québec, the Canadian Research Chair and the Université de Montréal.

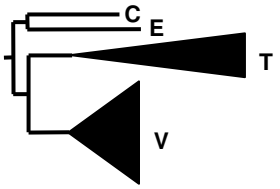
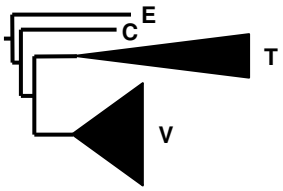
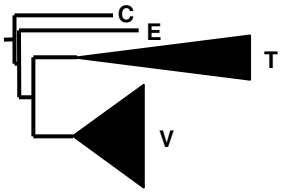
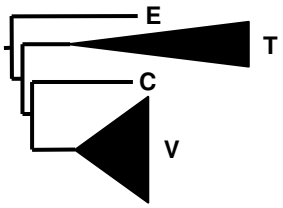
Authors' contribution H.P. conceived the study. D.C. contributed sequence data from the *Oikopleura* genome project. F.D., H.B. and H.P. assembled the dataset and performed phylogenetic analyses. F.D. wrote the first draft of the manuscript and all authors contributed to the writing of its final version.

Competing interests' statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.P. (herve.philippe@umontreal.ca).

Figure 1 | Phylogenetic analyses of genomic data strongly supports the grouping of tunicates and vertebrates into Olfactores. Maximum likelihood (ML) tree obtained from the analysis 33,800 aligned amino-acid positions under a WAG substitution matrix plus a four-category gamma rate correction ($\alpha = 0.5$) using two independent reconstruction algorithms (see Supplementary Information). Weighted maximum parsimony and Bayesian inference using the same WAG+F+Γ_4 model and WAG+F+Γ_4 plus covarion model also retrieved this same topology (see Supplementary Information). Bootstrap proportions obtained after 100 ML (red) and 1,000 MP replicates (blue), as well as Bayesian posterior probabilities (black) are shown for selected branches. A star indicates that all three values are maximal (100%, 100% and 1.0). Scale bar indicates number of changes per site.

Table 1 | Results of likelihood-based tests of alternative topologies within deuterostomes.

Trees	-Ln L	ΔLn L	AU	SH	RELL BP
	554,914.8	Best	0.947	1.000	0.938
	554,967.2	52.4	0.071	0.415	0.061
	555,051.5	136.7	0.000*	0.019*	0.000*
	555,066.4	151.6	0.004*	0.007*	0.002*

Results computed with a concatenated WAG+F+Γ_4 model are given for the top four ranking topologies, all other 11 alternative topologies being rejected by all tests with $p < 0.001$.

E: echinoderms, C: cephalochordates, T: tunicates and V: vertebrates.

* Statistically significant at the 5% level.

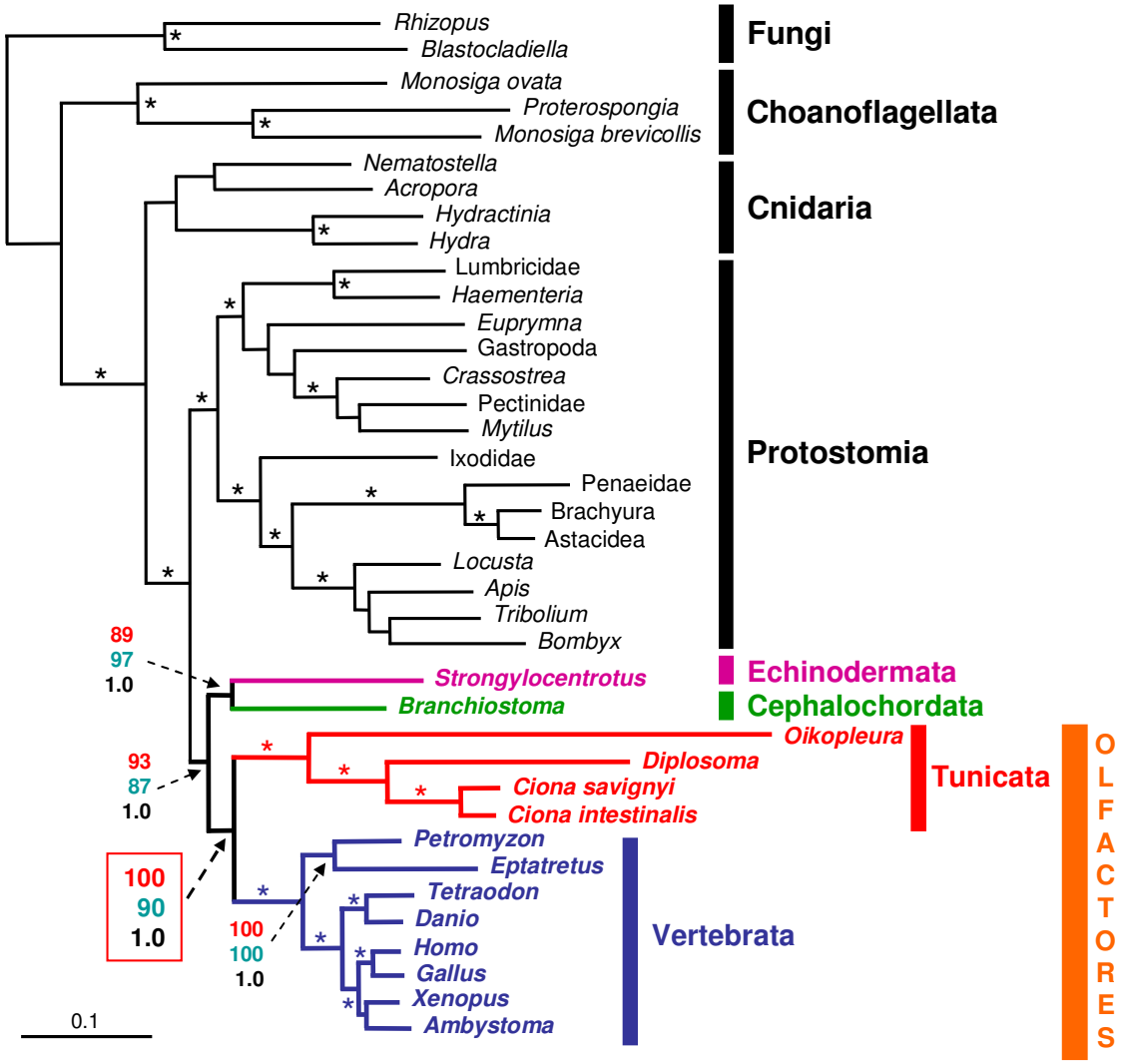

Rhizopus
Blastocladiella
Monosiga ovata
Proterospongia
Monosiga brevicollis
Nematostella
Acropora
Hydractinia
Hydra
Lumbricidae
Haementeria
Euprymna
Gastropoda
Crassostrea
Pectinidae
Mytilus
Ixodidae
Penaeidae
Brachyura
Astacidea
Locusta
Apis
Tribolium
Bombyx
Strongylocentrotus
Branchiostoma
Oikopleura
Diplosoma
Ciona savignyi
Ciona intestinalis
Petromyzon
Eptatretus
Tetraodon
Danio
Homo
Gallus
Xenopus
Ambystoma
0.1
89
97
1.0
93
87
1.0
100
90
1.0
100
100
1.0
Fungi
Choanoflagellata
Cnidaria
Protostomia
Echinodermata
Cephalochordata
Tunicata
Vertebrata
OLFACTORES

Figure 1

Tunicates and not cephalochordates are the closest living relatives of vertebrates

Frédéric Delsuc, Henner Brinkmann, Daniel Chourrout & Hervé Philippe

Supplementary Information

Supplementary methods

Data assembly

Each of the 146 gene alignments used in a previous study[1] were updated with newly available sequences downloaded from the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/) and the EST Database (http://www.ncbi.nlm.nih.gov/dbEST/) of GenBank at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) using new features of the program ED from the MUST package[2]. Unambiguous aligned regions were automatically detected and removed using the program GBlocks[3] and this selection was manually refined using the program ED. The list of genes with complete names and number of amino-acid positions are reported in Table S1.

The concatenation the 146 genes was then realised thanks to the program SCAFOS[4]. This program allows the selection of sequences according to their degree of divergence

using the ML distance matrix computed under a WAG+F model by TREE-PUZZLE[5]. It also permits optimising the percentage of missing data per taxa by creating chimerical sequences for species belonging to the same taxonomic group (see list below). The resulting alignment of 146 genes and 38 species for 33,800 unambiguously aligned positions is available upon request from HP.

List of chimerical Operational Taxonomic Units (OTUs)

To increase the amount of information, we created the following chimerical sequences between closely related taxa. The species the most represented in a given group is indicated in bold. Above the species level OTUs have been named from the inclusive taxonomic group containing the species:

Brachyura: ***Callinectes sapidus***, *Celuca pugilator*

Astacidea: ***Homarus americanus***, *Pacifastacus leniusculus*

Penaeidae: ***Litopenaeus setiferus***, *Litopenaeus vannamei*, *Penaeus monodon*, *Marsupenaeus japonicus*

Ixodidae: ***Amblyomma americanum***, *Amblyomma variegatum*, *Boophilus microplus*, *Rhipicephalus appendiculatus*

Gastropoda: *Aplysia californica*, ***Biomphalaria glabrata***, *Lymnaea stagnalis*

Pectinidae: ***Argopecten irradians***, *Chlamys farreri*

Crassostrea: *Crassostrea gigas*, ***Crassostrea virginica***

Lumbricidae: *Eisenia andrei*, ***Lumbricus rubellus***

Ambystoma: ***Ambystoma mexicanum***, *Ambystoma tigrinum*

Xenopus: ***Xenopus (Silurana) tropicalis***, *Xenopus laevis*

Eutheria: *Canis familiaris*, ***Homo sapiens***, *Pongo pygmaeus*, *Mus musculus*, *Rattus norvegicus*

Myxinidae: ***Eptatretus burgeri***, *Myxine glutinosa*

Branchiostoma: *Branchiostoma belcheri*, **Branchiostoma floridae**, *Branchiostoma lanceolatum*

Strongylocentrotidae: ***Strongylocentrotus purpuratus***, *Hemicentrotus pulcherrimus*, *Paracentrotus lividus*

The construction of a supermatrix containing a reasonable number of taxa unavoidably implied a certain amount of missing data. In our concatenated dataset the number of amino acid residues available for the most incomplete species is nevertheless already large with 6,175 positions for *Acropora millepora*. The complete dataset comprised 33,800 unambiguously aligned positions with a mean of 21,766 (64%) amino acid residues per taxa. These figures were even higher for deuterostomes with a mean of 27,216 (81%) amino acid residues per taxa, *Diplosoma listerianum* being the most incomplete species (8,788 positions) (Table S2). Under these conditions, the impact of missing data on phylogenetic inference can be considered as negligible (see Ref. [6-8]).

Phylogenetic analyses

Tree reconstruction

In order to check for concordance of the results, both PHYML[9] with either a BIONJ or a user-defined starting tree and TREEFINDER[10] were used to obtain ML trees under a concatenated model assuming the WAG amino acid substitution matrix[11], ML estimation of amino acid frequencies, plus a gamma distribution with four categories (WAG+F+Γ_4). ML bootstrap proportions were obtained after 100 pseudo-replicates generated with SEQBOOT[12].

Weighted MP heuristic searches were conducted using PAUP[13] with 10 random additions of species and TBR branch swapping. A stepmatrix computed from the PAM amino-acid substitution matrix[14] was used to allow taking into account the different amino-acid

substitution probabilities in a Maximum Parsimony framework. MP bootstrap percentages were obtained after 1,000 replications using the same heuristic search strategy using PAUP[13].

Bayesian inferences using the same WAG+F+Γ_4 model and a WAG+F+Γ_4 plus covarion model were performed with the parallel version of MRBAYES[15]. Each Bayesian analyses using 4 Metropolis-coupled Markov Chain Monte Carlo (MCMCMC) starting from a random tree and the program default prior probabilities on model parameters was repeated twice in order to control for an adequate mixing of the MCMCMC. Bayesian posterior probabilities were obtained from the majority rule consensus of the tree sampled after the initial burnin period as determined by checking the convergence of likelihood values across MCMCMC generations. 500,000 MCMCMC generations with sampling every 100 generations were ran under the WAG+F+Γ_4 model, whereas computational time constraints limited this number to 120,000 under the WAG+F+Γ_4+covarion model.

Taxon sampling

We tested the robustness of our results to long-branch attraction (LBA) by first varying the tunicate species sampling. We performed phylogenetic analyses on a reduced dataset where the fast evolving hagfish (*Eptatretus marinus*) was excluded to ensure that only slowly evolving vertebrates are considered, and where the very fast evolving *Oikopleura dioica* was chosen to represent tunicates. LBA is known to affect phylogenetic reconstruction when using maximum parsimony (MP)[16,17], whereas maximum likelihood (ML) is more robust[18-20]. As expected, the consideration of *O. dioica* alone resulted in an aberrant MP tree where this taxon emerged before cnidarians with 100% bootstrap support (Fig. S1). This result can obviously be interpreted as a long-branch attraction artefact causing *O. dioica* to be attracted towards the distantly related outgroup. By contrast, ML still supports (84%) a sister-group relationship between *O. dioica* and vertebrates (Fig. S2). Nevertheless, the aberrant MP topology was not rejected by the conservative SH test[21] (p = 0,35) computed using TREE-PUZZLE[5] under the WAG+F+Γ_4 model.

Second, we also tested the effect of removing the sea-urchin (*Strongylocentrotus*) from the complete dataset as an attempt to evaluate if the non-inclusion of ambulacrarians (echinoderms and hemichordates) affects the support observed for Olfactores. The ML likelihood tree obtained without including the echinoderm has the same topology (Fig. S4) as the ML tree obtained from the complete dataset (see Fig. 1). The only difference is a slight decrease in the bootstrap support for Olfactores (from 100% to 90%). This decrease can be interpreted as a probable LBA of the fast evolving tunicates towards the outgroup; the long branch of the outgroup is no longer broken by the sea-urchin, a phenomenon known to exacerbate the effect of LBA[17,22]. Such an interpretation is supported by the observation that the more sensitive MP method favours a tree where tunicates erroneously emerge before all other bilaterians with 66% bootstrap support (data not shown). Therefore, it is likely that the strong support observed for Olfactores with the complete dataset will not be significantly affected by the future addition of taxa belonging to the echinoderm side of the deuterostome tree such as hemichordates and xenoturbellarians.

Compositional bias

To explore the extend of compositional heterogeneity at the amino acid level, we performed a principal component analysis (PCA) of amino acid frequencies. The first three axes of the PCA explained 71% of the variance due to compositional differences among taxa with 46% for axis 1, 15% for axis 2, and 10% for axis 3. Projections of compositional vectors on the 1st axis have been plotted against those of the other two axes (Fig. S3). In accordance with the accelerated evolutionary rate of their genes, tunicates are also extreme with regards to their amino acid composition. However, this analysis revealed no obvious compositional bias that would potentially group tunicates and vertebrates, cephalochordates being much more similar in composition to vertebrates than tunicates. Also, cephalochordates and echinoderms do not seem to share any evident compositional bias that might explain their grouping.

Nevertheless, in order to verify that our results are not biased by heterogeneous amino acid compositions, we recoded our dataset into the six biochemical categories of Dayhoff[23], a protocol that has been efficient for a difficult, ancient, phylogenetic problem[24]. The MP analysis using PAUP[13] of this compositionally homogenised dataset slightly increased the MP boostrap support for most deuterostome nodes relative to standard MP (Fig. S5), lowering thus the probability of a compositional artefact. Moreover, the Bayesian analysis under a GTR+Γ_4 model using MRBAYES[25] of a more conservative 4-state coding (with C coded as missing data and MVIL and FYW pulled together) resulted in the same topology as in Fig 1 with posterior probabilities of 1.0 for all nodes (data not shown).

Heterotachy

To evaluate the effect of heterotachy on phylogenetic inference with our phylogenomic dataset, we performed three kinds of analyses.

First, ML analyses were conducted under a partitioned-likelihood model[26] allowing each of the 146 genes to have its independent branch lengths estimated under a WAG+F+Γ_4 model. Since an exhaustive search is not possible for the 38 taxa simultaneously, we generated using PROTML[27] all 10,395 possible topologies linking the 8 following groups: outgroup (fungi, choanoflagellates, cnidarians), arthropods, molluscs, annelids, echinoderms, cephalochordates, tunicates, and vertebrates. These topologies were then analysed under a partitioned JTT+F model using PROTML. The 2,000 best topologies were retained for the more time consuming search with a partitioned WAG+F+Γ_4 model. The likelihood of each tree for each gene and the corresponding branch lengths were computed using TREE-PUZZLE[5]. The likelihood of each position for each tree was then computed using CODEML[28] and the site-wise likelihood values were then used to compute the RELL bootstrap values of each topology based on 1,000 replicas. This model accounts for heterotachy among genes in the sense that each of the 146 genes is allowed to have its independent branch lengths. The

partitioned-likelihood analysis of the 2,000 best ranking topologies for each gene resulted in the same topology as in Fig. 1 (Fig. S6).

Second, we explored the effect of using a covarion model that can be considered as the modelling of a particular form of heterotachy where sites are switching from being free to vary (on) to being invariable (off)[29]. Bayesian analysis of the complete amino-acid dataset under a WAG+F+Γ_4 plus covarion model provided unambiguous support for the ML topology of Fig. 1, Bayesian support values being 1.0 for all nodes (Fig. S7).

Finally, we also used a home-made Bayesian mixture model handling heterotachy (Yan Zhou et al. in prep.) inspired from the one proposed by Kolaczkowski & Thornton[30] on a reduced dataset containing only the 14 deuterostomes. Sites where then sorted according to their posterior probability of being assigned to one of two heterotachous partitions, using a JTT+Γ_4 model and assuming either the Euchordates or the Olfactores topology. This resulted in two pairs of unequal datasets exhibiting heterogeneous branch lengths of 29,462 and 4,334 sites with the Euchordates topology and 29,220 and 4,576 sites with the Euchordates topology. Statistical tests of the three alternative topologies relating the four groups (echinoderms, cephalochordates, tunicates and vertebrates) were then performed on each of these fours datasets using TREE-PUZZLE[5] under a concatenated WAG+F+Γ_4 model. Both heterotachous partitions of each dataset pair recovered Olfactores as the best hypothesis, the other two alternative hypotheses being significantly rejected by the SH test at the 5% level (data not shown).

References

1. Philippe, H., Lartillot, N. & Brinkmann, H. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Mol. Biol. Evol.* **22**, 1246-1253 (2005).

2. Philippe, H. MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res.* **21**, 5264-5272 (1993).

3. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552 (2000).

4. Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. ScaFos: selection, concatenation and fusion of sequences. (in prep.).

5. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502-504 (2002).

6. Wiens, J. J. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst Biol* **52**, 528-38. (2003).

7. Philippe, H. et al. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol* **21**, 1740-52 (2004).

8. Wiens, J. J. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Syst. Biol.* (in press).

9. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).

10. Jobb, G., von Haeseler, A. & Strimmer, K. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.* **4**, 18 (2004).

11. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691-699 (2001).

12. Felsenstein, J. (Distributed by the author, Department of Genetics, University of Washington, Seattle, 2001).

13. Swofford, D. L. (Sinauer, Sunderland, MA, 2000).

14. Xu, W. & Miranker, D. P. A metric model of amino acid substitution. *Bioinformatics* **20**, 1214-1221 (2004).

15. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407-415 (2004).

16. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401-410 (1978).

17. Hendy, M. D. & Penny, D. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38**, 297-309 (1989).

18. Gaut, B. S. & Lewis, P. O. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* **12**, 152-162 (1995).

19. Huelsenbeck, J. P. Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17-48 (1995).

20. Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G. & Philippe, H. An empirical assessment of long branch attraction artifacts in phylogenomics. *Syst. Biol.* (accepted).

21. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114-1116 (1999).

22. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6**, 361-375 (2005).

23. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. in *Atlas of Protein Sequences and Structure* (ed. Dayhoff, M. O.) 345-352 (National Biomedical Research Foundation, Washington DC, 1978).

24. Hrdy, I. et al. *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **432**, 618-622 (2004).

25. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).

26. Yang, Z. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**, 587-596 (1996).

27. Adachi, J. & Hasegawa, M. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr.* **28**, 1-150 (1996).

28. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555-556 (1997).

29. Huelsenbeck, J. P. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* **19**, 698-707 (2002).

30. Kolaczkowski, B. & Thornton, J. W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980-4 (2004).

Supplementary Table S1 | List of the 146 gene names and number of amino-acid positions conserved for each gene alignment.

Abbreviated name	Complete gene name	Number of amino acid positions
ar21	Actin-related protein 2/3 complex subunit 3	127
arc20	Actin-related protein 2/3 complex subunit 4	163
arp23	Actin-related protein 2/3 complex subunit 1b	178
cct-A	T complex protein 1 alpha subunit	505
cct-B	T complex protein 1 beta subunit	484
cct-D	T complex protein 1 delta subunit	469
cct-E	T complex protein 1 epsilon subunit	491
cct-G	T complex protein 1 gamma subunit	443
cct-N	T complex protein 1 eta subunit	437
cct-T	T complex protein 1 theta subunit	360
cct-Z	T complex protein 1 ? subunit	476
cpn60-mt	Heat shock protein HSP 60kDa mitochondrial	477
crfg	Nucleolar GTP binding protein 1	356
ef2-EF2	Elongation factor EF2	741
ef2-U5	Elongation factor Tu family U5 snRNP specific protein	637
eif5a	Eukaryotic initiation factor 5a	119
fibri	Fibrillarin	210
fpps	Farnesyl pyrophosphate synthase	127
glcn	N-acetyl glucosamine phophotransferase	198
grc5	60S ribosomal protein L10 QM protein	206
hsp70-E	Heat shock 70kDa protein form E	503
hsp70-mt	Heat shock 70kDa protein, mitochondrial form	486
if1a	Eukaryotic translation initiation factor 1a	117

if2b	Eukaryotic translation initiation factor 2b	152
if2g	Eukaryotic translation initiation factor 2g	421
if2p	Eukaryotic translation initiation factor 2p	368
if6	Eukaryotic translation initiation factor 6	221
l12e-A	40S ribosomal Protein S12	103
l12e-B	High mobility group like nuclear protein 2 NHP2	82
l12e-C	High mobility group like nuclear protein 2 NHP2-like protein 1	116
l12e-D	60S ribosomal Protein L7a	218
mcm-A	minichromosome family maintenance protein 5	376
mcm-B	minichromosome family maintenance protein 2	401
metk	S-adenosyl-methionine synthetase	325
mra1	Ribosome biogenesis protein NEP1 C2F protein	152
nsf1-C	Vacuolar protein sorting factor 4b	241
nsf1-E	Vacuolar protein sorting factor, paraplegin-like protein	303
nsf1-G	26S proteasome AAA-ATPase regulatory subunit 8	253
nsf1-H	AAA-ATPase family protein CDC48-like protein	173
nsf1-I	putative 26S proteasome ATPase regulatory subunit 7	273
nsf1-J	26S proteasome AAA-ATPase regulatory subunit 6	352
nsf1-K	26S proteasome AAA-ATPase regulatory subunit 6a	251
nsf1-L	26S proteasome AAA-ATPase regulatory subunit 6b	258
nsf1-M	26S proteasome AAA-ATPase regulatory subunit 4	269
nsf2-A	Transitional endoplasmic reticulum ATPase TER ATPase	533
nsf2-F	Vesicular fusion protein nsf2	368
orf2	putative 28 kDa protein	160
pace4	protein chromosome 2 ORF 4	183

Table S1 continued

pace6	programmed cell death protein 5	62
psma-A	20S proteasome beta subunit macropain zeta chain	201
psma-B	20S proteasome alpha 1a chain	189

psma-C	20S proteasome alpha 1b chain	201
psma-D	20S proteasome alpha 2 chain	214
psma-E	20S proteasome alpha 1c chain	192
psma-F	20S proteasome alpha 3 chain	188
psma-G	20S proteasome alpha 6 chain	216
psma-H	20S proteasome alpha 1d chain	148
psma-I	20S proteasome alpha 1e chain	190
psma-J	20S proteasome alpha 1f chain	155
psmb-K	20S proteasome beta 7 chain	192
psmb-L	20S proteasome beta 6 chain	164
psmb-M	20S proteasome beta 5 chain	177
psmb-N	20S proteasome beta 4 chain	115
rad23	UV excision repair protein RAD23	130
rad51-A	DNA repair protein RAD51	303
rf1	Eukaryotic peptide chain release factor subunit 1	374
rla2-B	60S acidic ribosomal protein P1	66
rpl1	60S ribosomal Protein 1	211
rpl11b	60S ribosomal Protein 11b	168
rpl12b	60S ribosomal Protein 12b	157
rpl13	60S ribosomal Protein 13	135
rpl14a	60S ribosomal Protein 14a	98
rpl15a	60S ribosomal Protein 15a	204
rpl16b	60S ribosomal Protein 16b	162
rpl17	60S ribosomal Protein 17	164
rpl18	60S ribosomal Protein 18	180
rpl19a	60S ribosomal Protein 19a	180
rpl2	60S ribosomal Protein 2	248
rpl20	60S ribosomal Protein 20	148
rpl21	60S ribosomal Protein 21	149
rpl22	60S ribosomal Protein 22	84

rpl23a	60S ribosomal Protein 23a	131
rpl24-A	60S ribosomal Protein 24a	110
rpl24-B	60S ribosomal Protein 24b	121
rpl25	60S ribosomal Protein 25	117
rpl26	60S ribosomal Protein 26	119
rpl27	60S ribosomal Protein 27	131
rpl3	60S ribosomal Protein 3	372
rpl30	60S ribosomal Protein 30	101
rpl31	60S ribosomal Protein 31	105
rpl32	60S ribosomal Protein 32	122
rpl33a	60S ribosomal Protein 33a	94
rpl34	60S ribosomal Protein 34	108
rpl35	60S ribosomal Protein 35	116
rpl37a	60S ribosomal Protein 37a	81
rpl38	60S ribosomal Protein 38	64
rpl39	60S ribosomal Protein 39	51
rpl4	60S ribosomal Protein 4	104
rpl43b	60S ribosomal Protein 43b	91
rpl4B	60S ribosomal Protein 4b	280

Table S1 continued

rpl5	60S ribosomal Protein 5	261
rpl6	60S ribosomal Protein 6	107
rpl7-A	60S ribosomal Protein 7a	201
rpl9	60S ribosomal Protein 9	160
rpo-A	RNA polymerase alpha subunit	684
rpo-B	RNA polymerase beta subunit	1217
rpp0	60S acidic ribosomal protein P0 L10E	284
rps1	40S ribosomal Protein 1	236
rps10	40S ribosomal Protein 10	92

rps11	40S ribosomal Protein 11	137
rps13a	40S ribosomal Protein 13a	151
rps14	40S ribosomal Protein 14	135
rps15	40S ribosomal Protein 15	134
rps16	40S ribosomal Protein 16	137
rps17	40S ribosomal Protein 17	102
rps18	40S ribosomal Protein 18	152
rps19	40S ribosomal Protein 19	129
rps2	40S ribosomal Protein 2	208
rps20	40S ribosomal Protein 20	100
rps22a	40S ribosomal Protein 22a	130
rps23	40S ribosomal Protein 23	142
rps25	40S ribosomal Protein 25	90
rps26	40S ribosomal Protein 26	98
rps27	40S ribosomal Protein 27	82
rps28a	40S ribosomal Protein 28a	60
rps29	40S ribosomal Protein 29	54
rps3	40S ribosomal Protein 3	206
rps4	40S ribosomal Protein 4	255
rps5	40S ribosomal Protein 5	189
rps6	40S ribosomal Protein 6	205
rps8	40S ribosomal Protein 8	184
sap40	40S ribosomal protein SA 40kDa laminin receptor 1	195
sra	Signal recognition particle receptor alpha subunit SR alpha	201
srp54	Signal recognition particle 54 kDa protein	385
srs	Seryl tRNA synthetase	326
suca	Succinyl-CoA ligase alpha chain mitochondrial precursor?	276
tfiid	TATA box binding protein related factor 2	174
topo1	DNA topoisomerase I, mitochondrial precursor	362
vata	Vacuolar ATP synthase catalytic subunit A	527

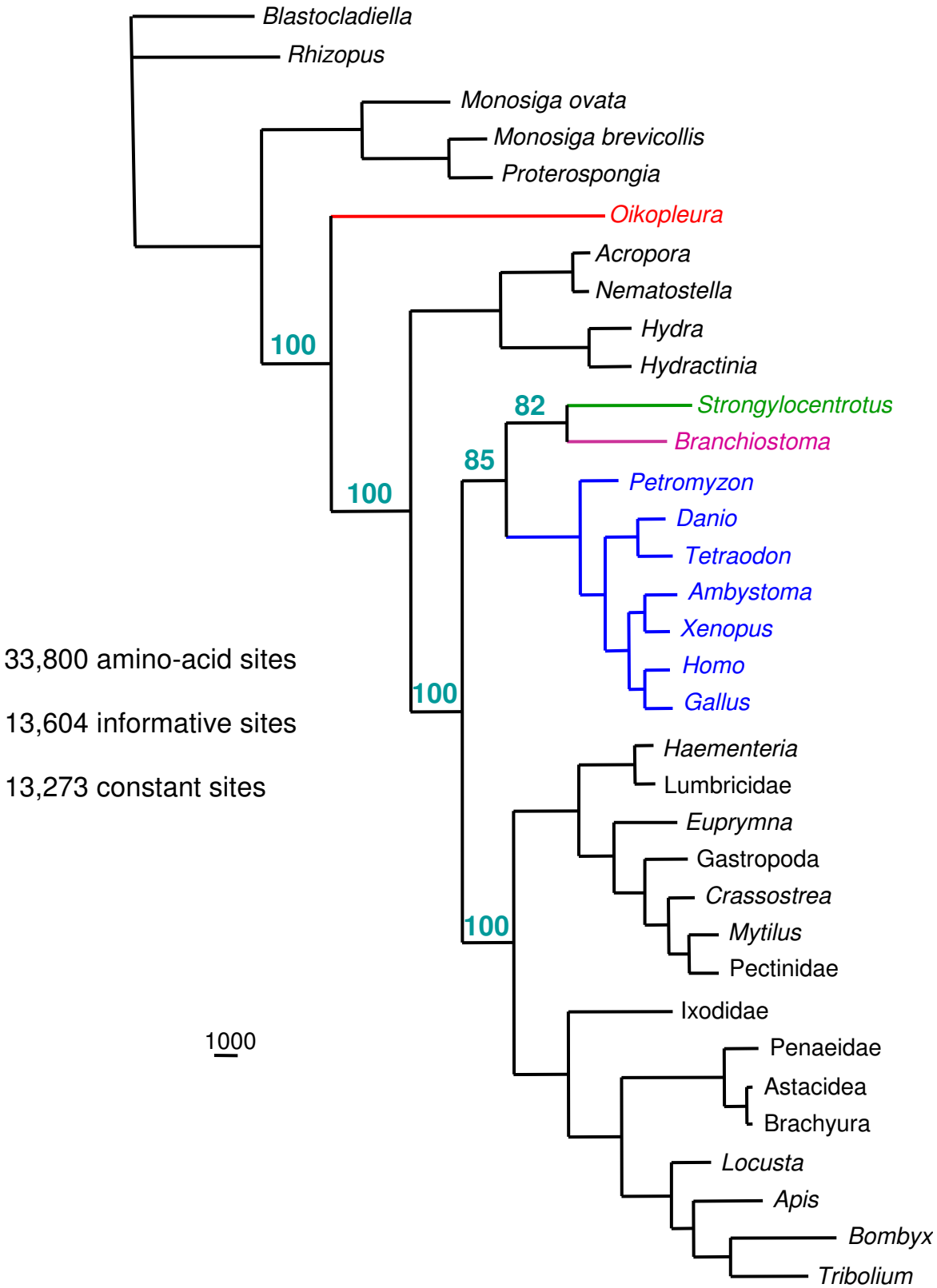
vatb	Vacuolar ATP synthase catalytic subunit B	451
vatc	Vacuolar ATP synthase catalytic subunit C	221
vate	Vacuolar ATP synthase catalytic subunit E	187
w09c	TGF beta inducible nuclear protein	248
wrs	tryptophanyl-tRNA synthetase	327
xpb	Helicase XPB subunit 2	450
yif1p	homolog of Yeast Golgi membrane protein	103

Supplementary Table S2 | Summary of the occurrence of missing data per taxa in the complete dataset.

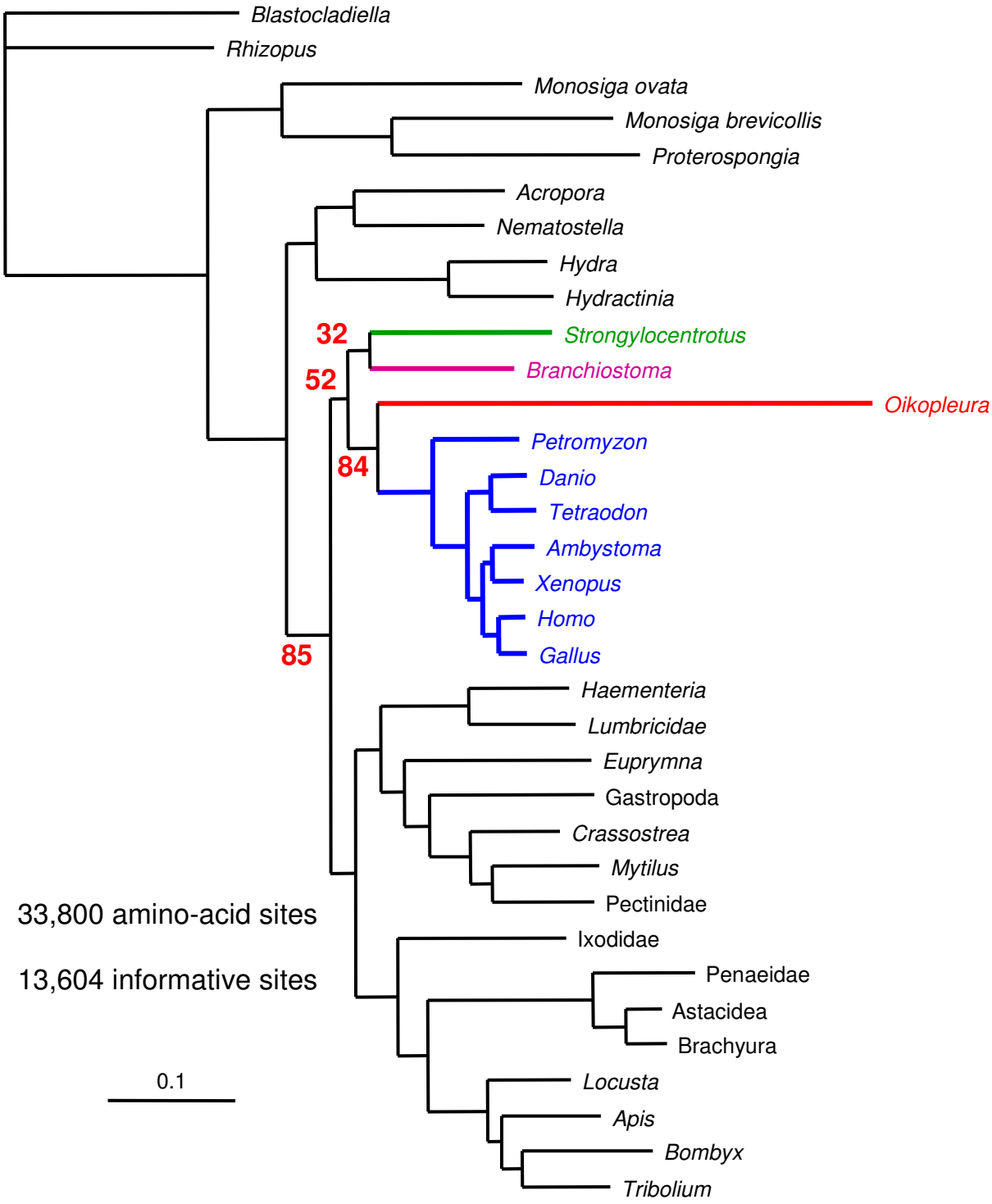
146 genes 33,800 amino acids	Occurrence in individual genes	Percentage of missing data	Number of amino acids
Acropora millepora	57	81.7	6,175
Haementeria depressa	54	80	6,755
Astacidea	69	75.7	8,215
Diplosoma listerianum	66	74	8,788
Proterospongia sp.	63	72.1	9,435
Mytilus galloprovincialis	78	66.5	11,318
Gastropoda	81	65.1	11,807
Pectinidae	79	64.4	12,048
Brachyura	67	64.3	12,059
Hydractinia echinata	96	57.2	14,467
Petromyzon marinus	98	50.9	16,607
Crassostrea	101	50.8	16,647
Penaeidae	110	50.7	16,662
Monosiga brevicollis	99	50.5	16,726
Lumbricidae	109	50.1	16,878
Euprymna scolopes	111	45.1	18,551
Locusta migratoria	123	44.7	18,698
Nematostella vectensis	137	36.9	21,339
Monosiga ovata	119	35.1	21,940
Myxinidae	124	32.4	22,863
Blastocladiella emersonii	134	26.7	24,765
Ambystoma	135	25.8	25,072
Branchiostoma	143	20.2	26,963
Ixodidae	134	18.8	27,463
Tetraodon nigroviridis	122	17.8	27,798
Apis mellifera	137	17	28,063
Hydra magnipapillata	138	14.9	28,780
Gallus gallus	127	14	29,073
Ciona savignyi	143	12	29,731
Bombyx mori	141	9.1	30,711
Oikopleura dioica	142	6.7	31,550
Danio rerio	144	5.7	31,878
Strongylocentrotidae	146	5.6	31,899
Ciona intestinalis	144	4.1	32,430
Xenopus	144	3.6	32,580
Rhizopus oryzae	145	1.8	33,208
Tribolium castaneum	145	1.3	33,357
Eutheria	146	0	33,794
Mean	**115**	**35.6**	**21,766**

Note: Deuterostomes are figured in red.

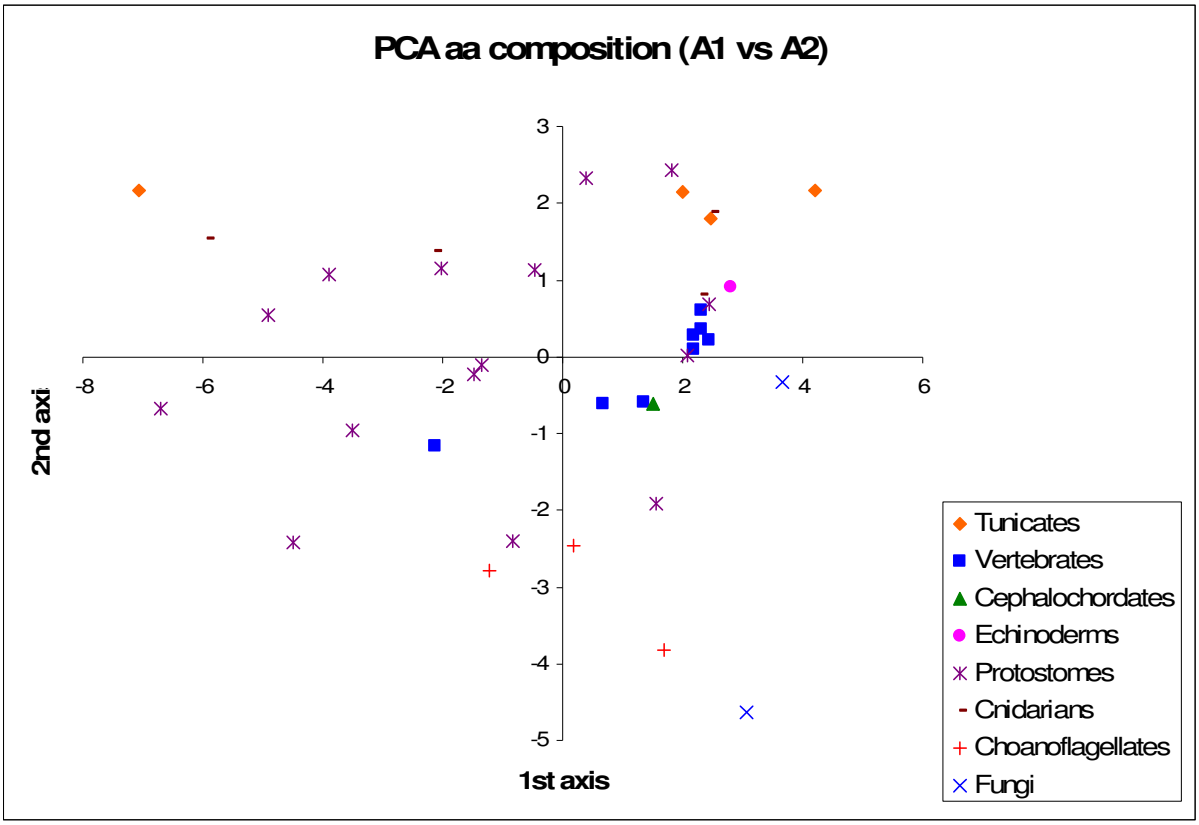
Supplementary Figure S1 | Most parsimonious tree obtained with a reduced dataset using *Oikopleura dioica* as the single representative of tunicates. Weighted MP heuristic searches were conducted using PAUP[13] with 10 random additions of species and TBR branch swapping and using a stepmatrix computed from the PAM amino-acid substitution matrix[14]. MP bootstrap percentages obtained after 1,000 replications with 10 random additions of species are shown for selected branches.

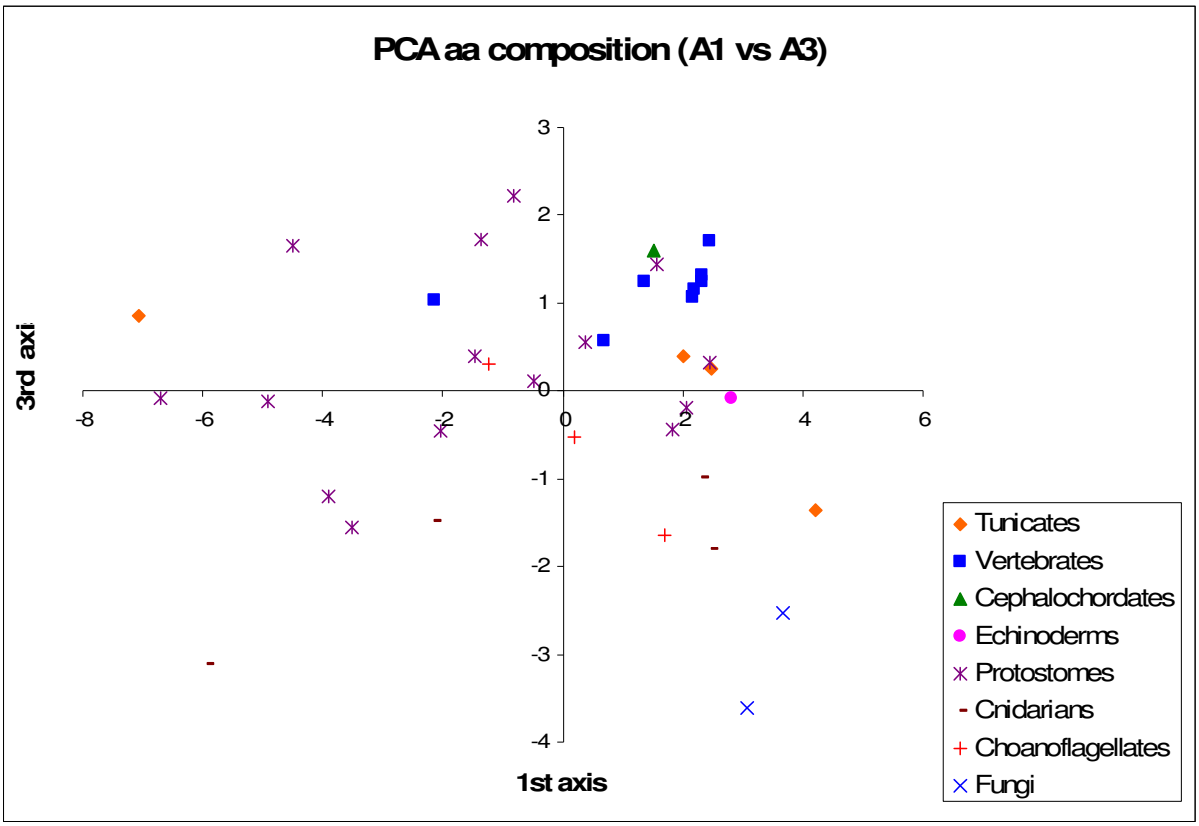

Blastocladiella
Rhizopus
Monosiga ovata
Monosiga brevicollis
Proterospongia
Oikopleura
Acropora
Nematostella
Hydra
Hydractinia
Strongylocentrotus
Branchiostoma
Petromyzon
Danio
Tetraodon
Ambystoma
Xenopus
Homo
Gallus
Haementeria
Lumbricidae
Euprymna
Gastropoda
Crassostrea
Mytilus
Pectinidae
Ixodidae
Penaeidae
Astacidea
Brachyura
Locusta
Apis
Bombyx
Tribolium
100
100
82
85
100
100
33,800 amino-acid sites
13,604 informative sites
13,273 constant sites
1000

Supplementary Figure S2 | Maximum likelihood tree obtained with a reduced dataset using *Oikopleura dioica* as the single representative of tunicates. This tree was inferred using PHYML[9] with a concatenated WAG+F+Γ_4 model using the tree inferred by MRBAYES[15] as a starting tree. Bootstrap values were computed from 100 replications starting from both a BIONJ tree (as usual) and the tree inferred by MRBAYES[15] in order to reduce the potential problem of local minima.

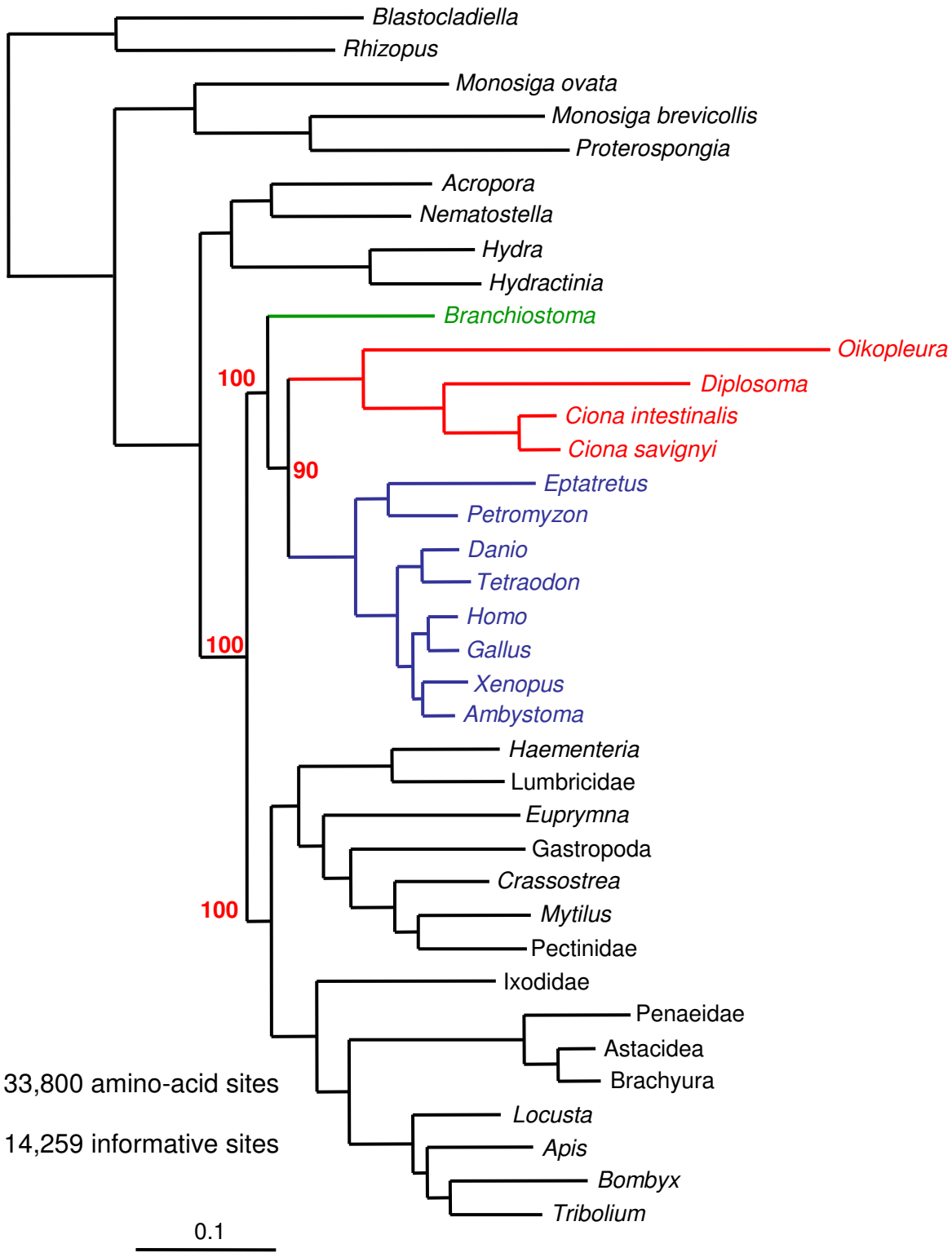

Blastocladiella
Rhizopus
Monosiga ovata
Monosiga brevicollis
Proterospongia
Acropora
Nematostella
Hydra
Hydractinia
32
Strongylocentrotus
52
Branchiostoma
Oikopleura
Petromyzon
84
Danio
Tetraodon
Ambystoma
Xenopus
Homo
Gallus
85
Haementeria
Lumbricidae
Euprymna
Gastropoda
Crassostrea
Mytilus
Pectinidae
33,800 amino-acid sites
Ixodidae
13,604 informative sites
Penaeidae
Astacidea
Brachyura
0.1
Locusta
Apis
Bombyx
Tribolium

Supplementary Figure S3 | Principal component analysis (PCA) of amino acid frequencies on the complete dataset.

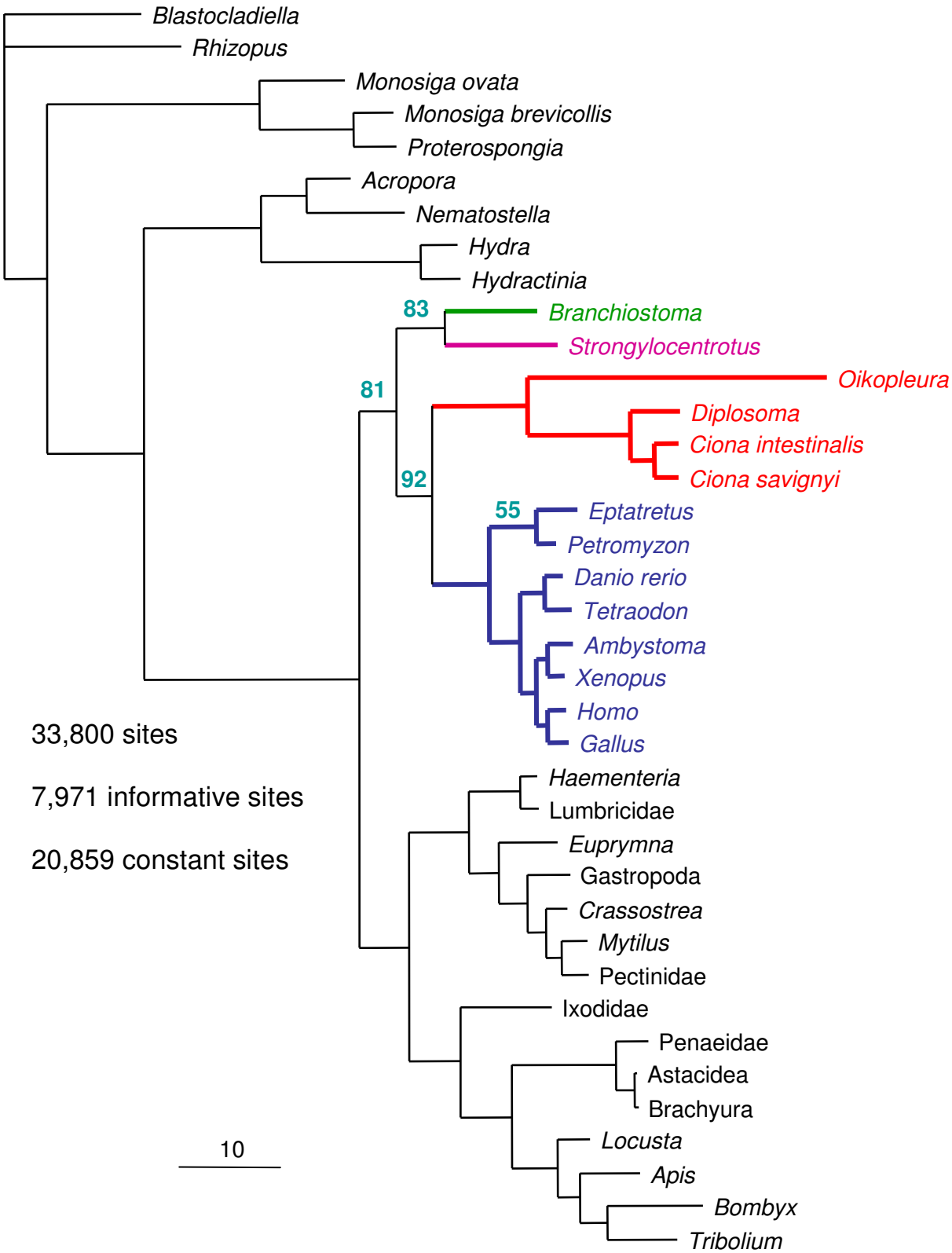



PCA aa composition (A1 vs A3)
3rd axi
1st axis
3
2
1
0
-1
-2
-3
-4
-8
-6
-4
-2
0
2
4
6
Tunicates
Vertebrates
Cephalochordates
Echinoderms
Protostomes
Cnidarians
Choanoflagellates
Fungi

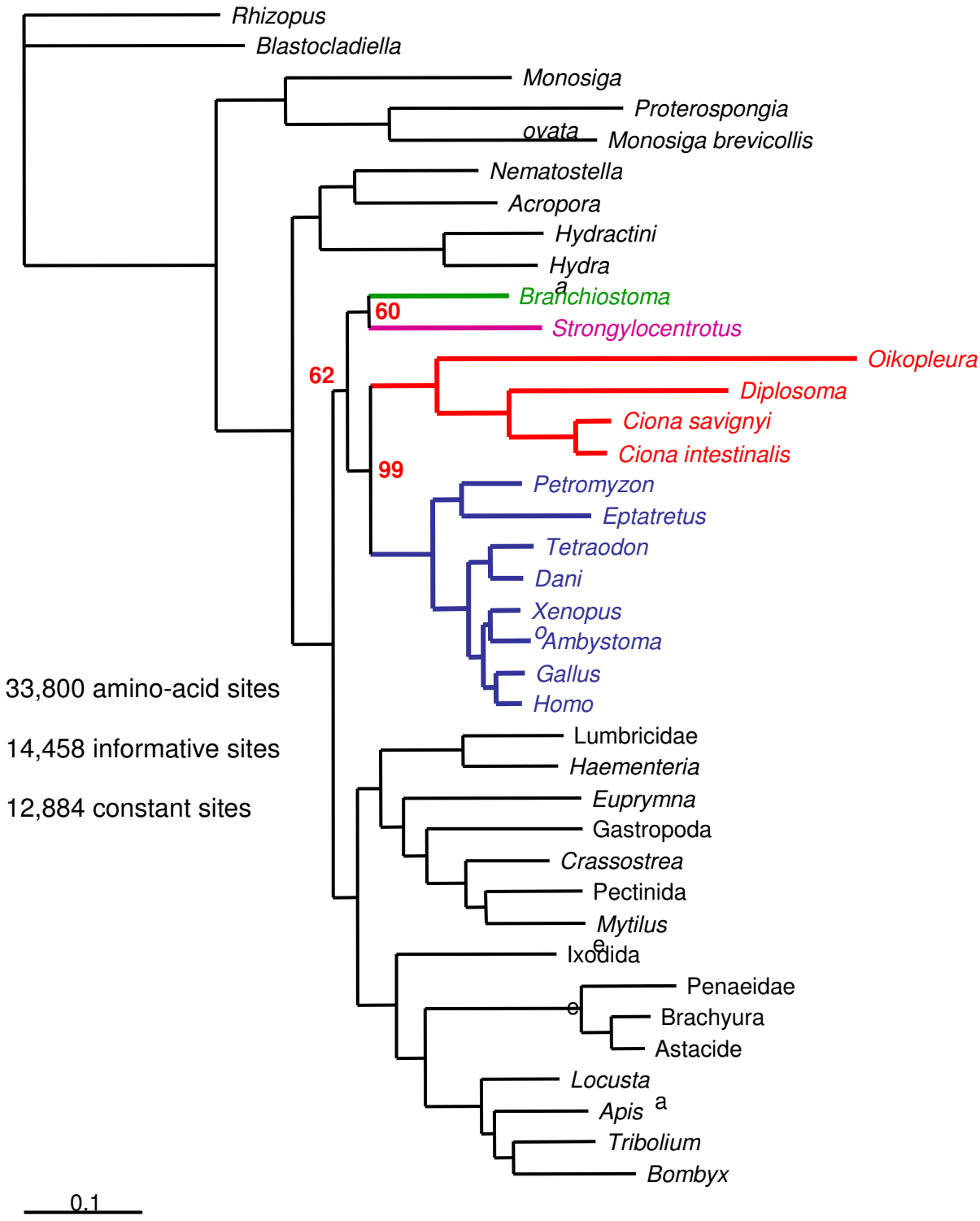
Supplementary Figure S4 | Maximum likelihood tree obtained after the removal of the sea-urchin (*Strongylocentrotus*) from the complete dataset. This tree was inferred using PHYML[9] with a concatenated WAG+F+Γ_4 model using a BIONJ starting tree. Bootstrap values were computed after ML 100 replications starting from the ML tree.


Blastocladiella
Rhizopus
Monosiga ovata
Monosiga brevicollis
Proterospongia
Acropora
Nematostella
Hydra
Hydractinia
Branchiostoma
Oikopleura
Diplosoma
Ciona intestinalis
Ciona savignyi
Eptatretus
Petromyzon
Danio
Tetraodon
Homo
Gallus
Xenopus
Ambystoma
Haementeria
Lumbricidae
Euprymna
Gastropoda
Crassostrea
Mytilus
Pectinidae
Ixodidae
Penaeidae
Astacidea
Brachyura
Locusta
Apis
Bombyx
Tribolium
100
90
100
100
33,800 amino-acid sites
14,259 informative sites
0.1

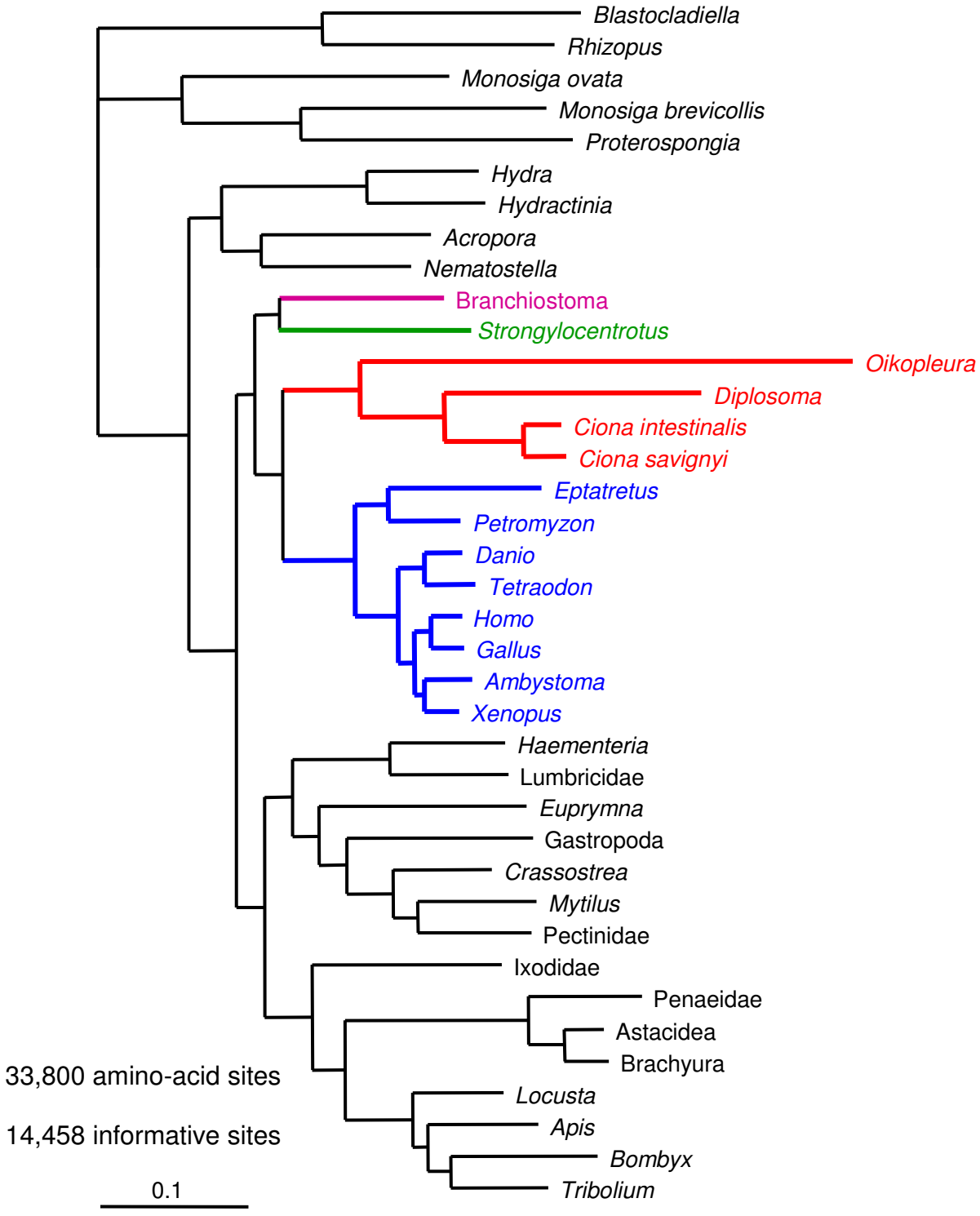
Supplementary Figure S5 | Most parsimonious tree obtained from the complete dataset recoded into six Dayhoff categories. MP heuristic searches were conducted using PAUP[13] with 100 random additions of species and TBR branch swapping. MP bootstrap percentages obtained after 1,000 replications with 10 random additions of species are shown for selected branches.


Blastocladiella
Rhizopus
Monosiga ovata
Monosiga brevicollis
Proterospongia
Acropora
Nematostella
Hydra
Hydractinia
83
Branchiostoma
Strongylocentrotus
81
Oikopleura
Diplosoma
Ciona intestinalis
Ciona savignyi
92
55
Eptatretus
Petromyzon
Danio rerio
Tetraodon
Ambystoma
Xenopus
Homo
Gallus
33,800 sites
7,971 informative sites
20,859 constant sites
Haementeria
Lumbricidae
Euprymna
Gastropoda
Crassostrea
Mytilus
Pectinidae
Ixodidae
Penaeidae
Astacidea
Brachyura
Locusta
Apis
Bombyx
Tribolium
10

Supplementary Figure S6 | Maximum likelihood topology identified by the partitioned-likelihood analysis on the complete dataset. Branch lengths were computed from the concatenated dataset using a WAG+F+Γ_4 model with TREE-PUZZLE[5]. RELL bootstrap values based on 1,000 replicates computed from the site-wise likelihoods obtained from CODEML[28] are shown for selected branches.


Rhizopus
Blastocladiella
Monosiga
Proterospongia
ovata
Monosiga brevicollis
Nematostella
Acropora
Hydractini
Hydra
Branchiostoma
60
Strongylocentrotus
Oikopleura
62
Diplosoma
Ciona savignyi
Ciona intestinalis
99
Petromyzon
Eptatretus
Tetraodon
Dani
Xenopus
Ambystoma
Gallus
Homo
33,800 amino-acid sites
14,458 informative sites
12,884 constant sites
Lumbricidae
Haementeria
Euprymna
Gastropoda
Crassostrea
Pectinida
Mytilus
Ixodida
Penaeidae
Brachyura
Astacide
Locusta
Apis
Tribolium
Bombyx
0.1

Supplementary Figure S7 | Majority rule consensus tree obtained from Bayesian analysis of the complete dataset under a WAG+F+Γ_4 plus covarion model. 4 MCMCMC were run in parallel[15] for 120,000 generations starting a random starting tree, sampling trees every 10 generations, and using the program default values for priors on model parameters. The consensus tree has been computed from the 10,000 trees sampled after the burnin period estimated to be 20,000 generations. Posterior probability values were maximal for all nodes (1.0).


Blastocladiella
Rhizopus
Monosiga ovata
Monosiga brevicollis
Proterospongia
Hydra
Hydractinia
Acropora
Nematostella
Branchiostoma
Strongylocentrotus
Oikopleura
Diplosoma
Ciona intestinalis
Ciona savignyi
Eptatretus
Petromyzon
Danio
Tetraodon
Homo
Gallus
Ambystoma
Xenopus
Haementeria
Lumbricidae
Euprymna
Gastropoda
Crassostrea
Mytilus
Pectinidae
Ixodidae
Penaeidae
Astacidea
Brachyura
Locusta
Apis
Bombyx
Tribolium
33,800 amino-acid sites
14,458 informative sites
0.1